Nunes Vieira, L. (2014). Indices of cognitive effort in machine translation post-editing. *Machine Translation*, *28*(3), 187-216. https://doi.org/10.1007/s10590-014-9156-x

[To cite this article, please use: Vieira, L.N. (2014). Indices of Cognitive Effort in Machine Translation Post-Editing. *Machine Translation* 28(3):187-216. doi: 10.1007/s10590-014-9156-x]

Indices of cognitive effort in machine translation post-editing

Lucas Nunes Vieira

Newcastle University
School of Modern Languages
Old Library Building
Newcastle upon Tyne, NE1 7RU
United Kingdom
e-mail: l.nunes-vieira@newcastle.ac.uk

Identifying indices of effort in post-editing of machine translation can have a number of applications, including estimating machine translation quality and calculating post-editors' pay rates. Both source-text and machine-output features as well as subjects' traits are investigated here in view of their impact on cognitive effort, which is measured with eye tracking and a subjective scale borrowed from the field of Educational Psychology. Data is analysed with mixed-effects models, and results indicate that the semantics-based automatic evaluation metric Meteor is significantly correlated with all measures of cognitive effort considered. Smaller effects are also observed for source-text linguistic features. Further insight is provided into the role of the source text in post-editing, with results suggesting that consulting the source text is only associated with how cognitively demanding the task is perceived in the case of those with a low level of proficiency in the source language. Subjects' working memory capacity was also taken into account and a relationship with post-editing productivity could be noticed. Scaled-up studies into the construct of working memory capacity and the use of eye tracking in models for quality estimation are suggested as future work.

Keywords: post-editing, cognitive effort, eye tracking, Meteor, working memory capacity.

1 Introduction

Previous large-scale experiments, carried out in both industrial and academic settings, indicate that, in comparison with traditional translation, post-editing (PE) is able to increase productivity and improve translation quality (see e.g. Green et al. 2013; Plitt and Masselot 2010). Despite its growing popularity, however, PE still involves a number of open challenges. These include, for example, the development of fair pricing schemes, as well as a reliable estimation of machine translation (MT) quality and PE effort. In view of this, being able to identify indices deemed to reflect how much effort PE is likely to pose has quickly become a topic of common interest in the field.

PE can be characterised as a process where attention is divided between three textual spheres: the source text (ST), the raw MT output, and the emerging target text (TT) (Krings 2001[1]). In this process, it would arguably be expected that characteristics from all these spheres as well as those stemming from individual traits may influence the amount of effort experienced by post-editors. In this respect, while the vast majority of previous research has addressed PE effort based on measures of time and/or amount of changes, the present paper reports on an attempt at predicting *cognitive* effort in PE, which is done based both on ST and MT-output characteristics, whilst also taking into account individual factors such as participants' working memory capacity (WMC) and source language (SL) proficiency.

[1] This is an English translation of work carried out in 1989-90.

Cognitive effort, as defined below, is measured with the use of eye tracking and a subjective scale proposed in the Educational Psychology literature. WMC is also analysed in view of PE productivity. The quality of the post-edited text – henceforth post-edited quality – is assessed with fluency and adequacy scales as well as with additional automatic evaluations. Eye-tracking data specific to the ST and TT is further analysed, and insights are also provided into the role of the ST in PE.

By analysing characteristics of both the ST and MT output, and taking into account the effort required to carry out the activity, PE is regarded here from a variety of angles, arguably allowing for a clearer picture of its cognitive implications. With regard to WMC, reading span – the measure of WMC adopted here – is considered in previous research (Temnikova 2010) as a ranking parameter for the PE difficulty of different MT errors, where problems stretching across longer spans are deemed harder to post-edit. Evidence from writing research also shows that WMC has a positive impact on monolingual reviewing, with skilled writers having larger memory spans and being more apt to implement global reviewing strategies (see McCutchen 1996:318). In view of these findings, it would arguably be expected that WMC also plays a role in PE, since post-editors capable of retaining longer spans in working memory may be more efficient in dealing with demanding MT errors. To the best knowledge of the present author, it is the first time that a measure of WMC is taken into account in a study on PE.

In view of the factors outlined above, it is hoped that the present study will serve to enhance the current understanding of the PE activity by pursuing three specific objectives: (i) identifying predictors of cognitive effort in PE, based on the ST, the MT output, and subjects' traits; (ii) investigating the role of the ST and SL proficiency in PE, drawing a parallel with how cognitively demanding participants perceive the task; and (iii) further investigating a potential relationship between WMC and PE productivity.

In the remainder of this paper, the theoretical background on the concept of cognitive effort and previous related work are outlined in section 2. The methodology adopted is explained in section 3. In section 4, results are reported and discussed. Conclusions are presented in section 5, where potential avenues for future work are suggested.

2 Background and related work

2.1 Defining and measuring cognitive effort

The concept of cognitive effort is known for its elusiveness. In the specific context of PE, Krings (2001) casts light on the issue by establishing a now recurrent three-fold definition of overall PE effort which involves a combination of *cognitive*, *technical* and *temporal* effort. Temporal effort consists of the time spent carrying out the task, and is deemed to directly reflect both cognitive and technical effort. Cognitive effort is defined by Krings (ibid. 179) as 'the type and extent of cognitive processes' involved in correcting/improving the MT output. Technical effort, in turn, is regarded as 'purely technical operations' (ibid. 179), being therefore distinguished from mental processing per se.

In the field of Educational Psychology, cognitive or mental effort is normally regarded as one of the components forming the overall notion of *cognitive load* – an overarching construct that also comprises the demands of the task, as well as aspects relating to subjects' performance (Paas 1992; DeStefano and LeFevre 2007; Paas and Van Merriënboer 1994). In this respect, terminology has been employed interchangeably in previous research, with cognitive *effort* and cognitive *load* being often used in allusion to the same constructs. In the Translation and PE literature, terms such as 'cognitive effort', 'cognitive load', and 'allocation of cognitive resources' can be observed.

Though previous research in Educational Psychology suggests that, under certain circumstances, cognitive effort alone can reflect overall cognitive load (Paas et al. 2003; Paas and Van Merriënboer 1994; Paas 1992; Hamilton 1979; Sanders 1979), cognitive effort is the term adopted here, referring specifically to 'the extent of cognitive processes' in PE (Krings 2001:179) or, more generally, to 'the amount of capacity or resources that is actually allocated to accommodate the task demands' (Paas and Van Merriënboer 1994:122) – a definition directly based on a notion of the concept initially laid out by Tyler et al. (1979:608).

As for the measurement of effort in PE, mounting evidence from previous research justifies the exploitation of cognitive effort in lieu of its temporal and technical counterparts. Koponen (2012) finds empirical evidence for a distinction between technical and perceived effort in PE by showing discrepancies

between ranks in a subjective scale of effort and the actual amount of changes carried out. She demonstrates that cognitive effort may be high even when few editing operations are performed. In addition, Koponen et al. (2012) investigate the PE process of sentences of similar length and which incurred similar technical effort, but discrepant PE durations. The authors show that different errors may lead to corrections that are equal in edit distance but different in PE time, providing further evidence of the pitfalls involved in taking only gross indicators of technical effort into account.

As regards purely temporal measures, in view of a lack of correlation between previous experience and total PE time, recent research suggests that 'PE effort and PE performance involve a high level of complexity that cannot be explained only by analysing temporal values' (De Almeida 2013:199). Similar results were obtained by Guerberof (2014), who found no significant incidence of experience on processing speed in PE.

In view of these studies, it could be argued that neither temporal nor technical parameters alone are robust for comprehensively analysing the PE process, which is why *cognitive* effort is tentatively taken into account here. Naturally, cognitive effort cannot be measured directly. This has motivated the use of a number of different measures in previous research, some more reliable than others. Krings (2001) uses think-aloud protocols as a tool, a method shown to interfere with cognitive processes in Translation (see e.g. O'Brien 2005; Jakobsen 2003). A ratio of PE time spent on pauses did not vary significantly between source sentences deemed more and less susceptible to being translated correctly by MT in O'Brien (2006a), who suggests pauses on their own may not be a robust measure for estimating cognitive effort in PE.

More recently, positive results have been obtained for improved versions of some of these measures. Koponen et al. (2012) demonstrate that average seconds per word can be used for estimating cognitive effort in PE. In addition, averaged pause measures showed a correlation with a ratio of complete editing events per number of words (EWR) in Lacruz and Shreve (2014), who use EWR as their primary indicator of cognitive effort.

In the present study, physiological and subjective parameters already traditional in the neighbouring fields of Cognitive and Educational Psychology are exploited. Room appears to exist for future research into how traditional measures in these fields correlate with findings put forth by Koponen et al. (2012) and Lacruz and Shreve (2014).

2.2 Source-text characteristics and post-editing effort

At an early stage, Krings (2001) shows that STs deemed more difficult lead to a higher amount of think-aloud verbalisations and a higher frequency of certain cognitive PE processes.

Bernth and Gdaniec (2002) focus on the modification of the ST as a way of guaranteeing an output of higher quality. The authors suggest a number of ways in which the machine translatability of the ST, i.e. how MT-friendly it is, can be enhanced by avoiding features such as ambiguity, ellipsis, etc. Features of this kind are usually referred to as negative translatability indicators (NTIs) (Underwood and Jongejan 2001) and have been particularly useful in the development of controlled languages (CL), which are usually more susceptible to being translated correctly by MT systems. In this respect, a number of previous studies looked at the relationship between controlled input and PE (Aikawa et al. 2007; O'Brien 2006b; Temnikova 2010), where good results are usually obtained for the use of CL.

O'Brien (2004) analysed the relationship between NTIs and PE effort and found that, on average, the presence of NTIs in the ST increases PE effort. The author highlights, however, that different NTIs will have different levels of impact. Abbreviations and proper nouns, for example, were amongst the NTIs for which no impact was observed.

Tatsumi (2009) makes one of the first attempts at statistically predicting PE time and shows that, in addition to automatic evaluation metrics (AEMs) used as an index of amount of edits, sentence-complexity features were necessary to increase the fit of regression models used in the analysis. Tatsumi (2010) also shows that a ST complexity score provided by the MT system SYSTRAN correlated well with PE time. In regard to findings in Tatsumi (2010), Green et al. (2013) point out that the statistical approach used fail to more solidly inform the matter of effort prediction in PE because the regression technique implemented does not account for variation between participants in the study and the wider population. In accounting for such variation, mixed-effects models have been considerably favoured in the literature (Balling 2008; Baayen et al. 2008; Green et al. 2013).

Green et al. (2013) have parsed English STs with Stanford CoreNLP[2] so as to obtain information on ST linguistic effort predictors. Results showed that the proportion of nouns in the ST was significantly correlated with PE time, with adjectives also being frequently hovered over with the mouse.

In Aziz et al. (2014), sentences as well as production units (PU) are considered for an analysis of ST linguistic features. PUs consist of clusters of editing operations that can be considered to form one complete unit (Carl and Kay 2011). Results indicate that PUs involving verbs tended to be associated with more PE time overall, while PUs involving nouns tended to be associated with more edits. Similarly to Green et al. (2013), however, this study does not take potential measures of MT quality into account. In addition, English is the SL used, when it cannot be excluded that ST linguistic features may not generalise across different languages. In the present study, non-English input is analysed and MT output evaluation is also performed – a contrast not normally addressed in previous research, especially whilst having cognitive effort as a response variable.

2.3 Machine translation evaluation and post-editing effort

As regards MT output quality and its impact on effort, early work in the field was also carried out by Krings (2001). He collected human ratings on MT quality at a sentence level and observed that, contrary to expectations, sentences of medium quality were the ones that posed the highest level of effort, postulating that these sentences required 'a greater dispersion of attention across three different texts' (ibid: 540), i.e. ST, raw MT and emerging TT.

Based on automatic MT evaluation, O'Brien (2011) shows a linear relationship between eye-tracking measures and AEMs previously tested in Tatsumi (2010), but final quality and ST-complexity variables are not measured – aspects that are addressed in the present study.

Features that are able to predict PE effort have also been a target of interest in the field of MT quality estimation (QE), which aims at producing automatic quality evaluation scores that dispense with human reference translations. These initiatives are closely related to the present study in that they usually involve the use of textual predictors of PE effort – effort in this context being normally regarded as a proxy for MT quality.

Early work in this field involves the automatic estimation of the machine translatability of the ST as a way of predicting output quality. Blatz et al. (2004) show the value of using predictors based both on the ST and the MT output. Among the features used are source-sentence length, frequency of n-grams – i.e. contiguous sequences of words –, language model (LM) probabilities, as well as correspondence probabilities between ST and MT output.

In Specia et al. (2009) and Specia et al. (2010), subjective PE effort annotations are exploited as a way of estimating MT quality. Specia (2011) also measures temporal and technical effort and results of a task-based test show an increase in PE productivity for sentences with high QE scores, with best results being obtained for sentences scored with QE systems based on time, as opposed to subjective annotations.[3] As regards the features tested, while for English STs percentage of nouns is amongst the indicators presenting a high correlation with editing time, features with highest correlation for French are mostly based on LM probabilities. Results in Specia (2011), however, are based on a single translator per language pair.

The studies mentioned above share similar goals pertaining to the identification of textual characteristics deemed to predict PE effort. With a few exceptions, the approach adopted in most of these studies is based on PE time or amount of changes, taking either ST complexity or MT-output quality into account as predictors. The present study, by contrast, aims to focus mainly on cognitive effort as a parameter. In addition, non-English input is adopted here and mixed-effects models are used for statistical analysis, tentatively tackling the already known research issue posed by subject variation in PE, an aspect not statistically handled in the majority of previous research. With regard to this variation, participants' WMC and SL proficiency are also measured, allowing for insights into individual characteristics accounting for cognitive effort and productivity in PE.

As regards the field of QE, by making use of features normally included in QE systems, results reported here are also able to provide further insight into the cognitive implications of these features, potentially leading to informed decisions in the estimation of MT quality for the French-English language pair.

---

[2] http://nlp.stanford.edu/software/corenlp.shtml

[3] More recently, QE has been handled in the context of the QT Launchpad project (http://www.qt21.eu/launchpad/), and annual shared tasks within the Workshop on Statistical Machine Translation (WMT) (http://www.statmt.org/).

3 Methodology and procedure

3.1 Selection of source texts

Excerpts of two news articles originally published in French were selected for the present study from the *newstest2013* dataset, which results from the 2013 edition of the Workshop on Statistical Machine Translation (WMT).[4] The selected texts are about prostate cancer screening[5] (text A) and the voting system in the United States (text B).[6] Care has been taken in attempting to choose texts that had different levels of translating difficulty and that could generate plausible translations in English – and which also did not have marked culture-specific items or concepts only understandable to a given group of readers.

To assess potential translating difficulty, each text was divided into 3 passages of at least 100 words, respecting paragraph and sentence boundaries. The lexical frequency and non-literalness of these passages was then measured, and they were also assessed with readability formulae that can be used for French – namely Kandel and Moles (KM) (Kandel and Moles 1958), and LIX[7] formula (Bjornsson 1968). In previous research (Hvelplund 2011; Jensen 2009) these measures are considered 'strong indicators of source-text difficulty in translation' (Hvelplund 2011:88), which motivates their use here. Similarly, nouns and adjectives are reported in Green et al. (2013) as showing an association with PE time and mouse hover patterns for English. In that way, the combined percentage of nouns and adjectives in the passages is also taken into account, further testing the impact of these part-of-speech (POS) categories for French.

Lexical frequency was calculated based on the percentage of words in the passages that could be found on a list of the 1000 ('1K') most frequent words in French (Jones 2000).[8] POS percentages were obtained by parsing the texts with Stanford French Parser (Green et al. 2011) and extracting POS features with regular expressions. Passages 1-3, in text A, have 'easier' figures overall, while passages 4-6, in text B, were assessed as being 'more difficult'. In the LIX scale, higher values indicate less readability, while higher KM values indicate higher readability – details are presented in Table 1.

Table 1 Information on the STs selected for the study - shaded rows have total values for each text (figures exclude titles and introductory paragraphs, which were not considered for data analysis)

	Freq. (% on 1K list)	LIX readability	KM readability	Non-lit. expression count	%N+Adj	Word Count	Avg. sentence length (in chars, w/ white spaces)
Passage1	79.3%	45.6	58	3	32%	109	118±73 SD
Passage2	84.1%	47.8	68	3	27%	150	136±96 SD
Passage3	82.3%	46.2	61	2	21%	135	124±41 SD
Text A (easy)	82.2%	47.8	63	8	26%	394	122±68 SD
Passage4	75.2%	56.3	48	9	44%	172	131±35 SD
Passage5	75.8%	54.2	50	1	40%	134	126±52 SD
Passage6	81.5%	55	54	2	34%	144	152±46 SD
Text B (diff.)	77.4%	55	51	12	39%	450	136±44 SD

It should be noted that this prior assessment serves mainly as a way of making sure that participants are exposed to texts expected to pose different levels of translating difficulty. The actual impact of ST features on cognitive effort is analysed at a sentence level in the present study (see section 3.7). Out of all news articles in the *newstest2013* dataset that fit the criteria of not having culturally marked items or overly specific subject

[4] http://www.statmt.org/wmt13/

[5] Original available at http://www.lapresse.ca/vivre/sante/201211/30/01-4599309-depistage-du-cancer-de-la-prostate-passer-le-test-ou-non.php

[6] Original available at http://www.lapresse.ca/la-tribune/opinions/201207/30/01-4560667-une-strategie-republicaine-pour-contrer-la-reelection-dobama.php

[7] Abbreviation for *läsbarhetsindex* (readability index), in Swedish.

[8] See http://www.lextutor.ca

matters, these were the texts with the most discrepant percentages of nouns and adjectives combined. Other measures roughly follow this pattern, as can be observed in Table 1.

By collating the source texts in the *newstest2013* dataset with the original articles published online it was noted that short introductory paragraphs after the title and, in one of the texts, two segments in the body of the article, could not be found in the dataset versions. To make sure participants did not miss potentially relevant information, sentences in these additional passages were included in the materials, and the texts were presented for editing starting from the title and in their original order. Translations for these extra sentences were randomly selected from online and commercial MT systems, but data produced with them has not been analysed because (i) reference translations were not available for these sentences, and (ii) even though a warm-up task was carried out, it was deemed desirable to disregard data produced with sentences at the very beginning of the excerpts to avoid acclimatisation effects. Overall, 41 sentences and 844 source words were considered for analysis.

3.2 Selection of machine translation output

The *newstest2013* dataset was also used for the selection of the MT output. To increase variability in quality in the sample, in addition to anonymised outputs from online systems already included in the dataset, translations produced with a further three online and two commercial systems were selected[9], composing a corpus of 24 candidate translations (19 already in the dataset plus five newly harvested ones) for each source sentence. By making use of the *newstest2013* dataset, the study includes translations from either online/commercial engines or statistical systems tuned with WMT news data, the text genre adopted for the investigation.

Version 1.4 of the Meteor metric (Denkowski and Lavie 2011), run at default settings, was taken into account for the selection. Meteor measures the similarity between a hypothesis-translation and a reference. It differs from more traditional AEMs in that it takes semantic information into account, such as synonymy and paraphrasing, as well as stemming. Meteor scores vary from 0.0 to 1.0, where 1.0 represents the perfect match of a hypothesis with a reference. Illustrative examples of Meteor scoring are provided below.

Ref: In addition, five million new voters in 2012 do not have such identification.
Hyp 1: *In addition*, five million new voters in 2012 do not have such identification. (1.0)
Hyp 2: *What is more*, five million new voters in 2012 do not have such identification. (0.95)
Hyp 3: *In contrast*, five million new voters in 2012 do not have such identification. (0.53)

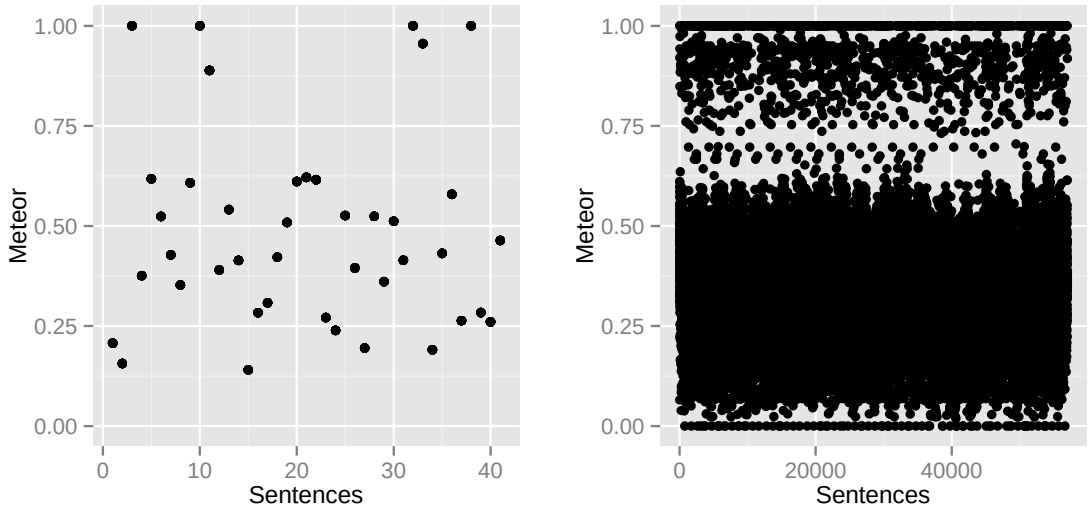
As can be seen above, Hyp 1 is a perfect match with the reference, thus receiving the Meteor score of 1. Hyp 2, though bearing a different term to the reference – 'what is more' instead of 'in addition' – receives a high score (0.95), which can be explained by their semantic proximity. Hyp 3, in turn, modifies the meaning of the reference by the use of 'in contrast'. Hyp 3 receives the comparatively lower score of 0.53. In view of this additional functionality, it could be argued that semantics-based AEMs can be more sensitive to MT output accuracy, which might not be the case for traditional AEMs based only on surface-level similarity.

The human reference translations used here are those included in the *newstest2013* dataset. For the selection of machine translations, Meteor was considered in terms of ranges in the space of one decimal place, with scores from 0.0 to 0.09 (inclusive) representing the first (bottom) range, and scores from 0.9 to 1.0 (inclusive) representing the tenth (top) range. Of ten possible ranges, 0.0-0.09 and 0.7-0.79 were absent from the materials. Translations in these ranges, however, account for approximately 1% of all 57K sentences in the FR-EN *newstest2013* dataset. In that way, the absence of these scores from the study was not deemed a sampling problem.

Of the 24 potential translations for each source sentence, at a first step one candidate per Meteor range was randomly selected. To eliminate overlapping versions, the distribution of the score across all sentences was taken into account. Translations were randomly selected from the least to the most represented Meteor range, aiming for the best possible balance of score ranges in the sample. The distribution of the resulting selection can be observed in Fig. 1.

[9] The systems used were SDL Freetranslation.com (`http://www.freetranslation.com`), PROMT (`http://www.online-translator.com/?External=aspForms&prmtlang=en`), TransPerfect (`http://web.transperfect.com/free-translations/`) Microsoft Translator, via MS Word, and SDL Automated Translation, via Trados Studio 2011(all harvested in October 2013).

Fig. 1 Meteor distributions in study sample (left) and in the FR-EN *newstest2013* dataset (right)



As can be seen, the distribution of translations in the sample is comparable to the distribution of the entire FR-EN *newstest2013* dataset. By adopting such a comprehensive approach, translations with scores as low as 0.14 were present in the materials. Upon visual inspection it could be noted that translations below 0.20 were of extremely poor quality, including non-translated and malformed words.[10] In that respect, a decision had to be made as to whether to keep these sentences or clean the sample. Since Meteor is one of the features being tested, having the entire spectrum of the metric seemed desirable. In view of this, these sentences were kept in the first instance, but further tests were carried out by excluding machine translations with low human-assessed scores to check if their potentially outlying impact had an influence on results (see section 4.1).

3.3 Participants

Recent PE research suggests that there is a complex relationship between experience (in either PE or Translating) and PE effort (see section 2.1), and that aspects such as adherence to guidelines can have a more visible association with PE performance (De Almeida 2013). In addition, gains in translation fluency have been observed even when PE is carried out by non-professionals without access to the ST (Mitchell et al. 2013). In view of these findings, it seemed interesting to allow for a varied sample in the present study. Since the impact of SL proficiency on cognitive effort is also an interest, French proficiency also varies between participants.

Potential subjects received an information sheet with details regarding the general purpose of the research and the practicalities involved in taking part. The decision to participate was voluntary and made upon awareness of the details on this sheet, such as the fact they would have their eye movements recorded and their WMC and French proficiency measured. Prior ethical approval was obtained from the ethics committee at the author's institution for the recruitment of human participants for the study.

Fourteen subjects were recruited overall, of whom 13 produced usable data.[11] Two subjects were postgraduate Translation students (PG) who had previous relevant experience as translators and translation revisers. Six were final-year undergraduate students (UG) of Modern Languages, of whom one had previous experience as an in-house translator. Two were established professionals who had at least three years' experience (P), of whom one had experience in PE. Three were non-professionals who had little to no translating experience (NP). All subjects who were not established professionals already held, or were in the process of obtaining, a higher education degree in Translation or related area involving the study of translation. All participants were native speakers of English. Further details on their profile, including level of French proficiency and WMC, are presented in Table 2.

[10] One of the systems seemed to have a preprocessing problem and certain words with French accented characters were not properly handled. This, however, was not a consistent pattern.

[11] Accidental incorrect usage of the editing interface rendered data from one participant unreliable.

French proficiency was measured with a yes/no vocabulary recognition task including plausible non-words (Meara and Buxton 1987).[12] Tasks of this kind have been tested in a number of previous studies and are normally deemed robust 'particularly for placement and diagnostic purposes' (Read 2007:113).

Table 2 Participants' profile. PG – postgraduate students; UG – undergraduate students; NP – non-professionals; P – established professionals

	P01	P02	P03	P04	P05	P06	P07	P08	P09	P10	P11	P12	P13
Experience (in years)	1.5	-	-	-	-	< 1	< 1	-	-	4	-	13	3
FR Vocab Test (0-100)	79	95	13	88	95	50	97	95	97	89	18	60	95
Absolute Reading Span (0-75)	25	42	56	14	36	56	44	68	61	38	61	32	37
Education/Status	PG	UG	UG	NP	NP	NP	UG	UG	UG	PG	UG	P	P
Age	23	21	22	25	55	27	22	21	22	26	21	40	60

Participants' WMC was measured with the automated version of a reading span task[13] (Unsworth et al. 2005; Unsworth et al. 2009). This task too has been subjected to previous tests (Redick et al. 2012). For economy of space, only absolute WMC scores are presented in Table 2, but both absolute and partial scores have been tested in the analysis.

3.4 Conducting the task

Prior to the PE task, a warm-up trial was conducted to acquaint participants with the set-up. PET (Aziz et al. 2012) was the editing platform adopted. Though document order was maintained, participants could only see one sentence at a time.[14] Similarly, they were not allowed to revisit sentences – a level of control that was necessary to guarantee a reliable collection of eye-tracking data.

Participants were asked to consider a one-page editing brief with instructions on the level of PE and quality expected. Parameters for the upper quality band suggested in the TAUS MT PE guidelines (TAUS 2010) were included in the brief. The task was carried out with no strict time pressure, but participants were instructed to keep as much of the existing MT output as possible, doing so as fast as they could, without over-pondering on decisions. After confirming each sentence, they were automatically prompted to rate perceived cognitive effort on a scale (see section 3.5) configured within PET's interface. While the use of dictionaries or Internet searches was not allowed, before editing each text all participants were asked to read a short document with background information on the respective text's topic. PET's interface, as set-up for the task, can be observed in Fig. 2.

Fig. 2 Set-up of PET interface – ST on the left and TT on the right

[12] See http://www.lextutor.ca/

[13] The WMC test was carried out twice for P04, as processing times longer than the participant's normal values (automatically calculated in the training phase of the test) could be seen as invalidating data from the first run. Using the second score does not influence the results reported and is not seen as problematic here, as both scores were low in comparison with those obtained by other participants.

[14] Two passages in text A had quotes that trespassed sentence boundaries. In view of the interface adopted, opening and closing inverted commas were added to each separate sentence in these cases.

As regards the sequence of tasks to be carried out, WMC and vocabulary tests were taken after post-editing the documents in an attempt to avoid an impact of fatigue on the data deemed central for the present study – i.e. data reflecting cognitive effort in PE. The order of presentation of the texts (A or B) was counterbalanced to avoid fatigue effects exerted by the PE task itself. A break was allowed in between texts. After carrying out the WMC task, participants were also asked to provide information on prior experience and level of education by filling out an online questionnaire, which included the French vocabulary test. The average time taken to post-edit both texts was 43 min (±13 min SD). The WMC task takes approximately 20 min to complete, and the questionnaire, including the vocabulary test, took 10 min to fill out, on average.

3.5 Measuring cognitive effort

Average fixation duration, fixation count and a self-report scale were used to approximate cognitive effort. In the context of eye-movement behaviour, a fixation is defined as 'the state when the eye remains still over a period of time' (Holmqvist et al. 2011:21) and their count and average duration have been largely used as measures of cognitive processing in reading research – see Rayner (1998) for an extensive review –, having been previously used also in PE (e.g. O'Brien 2011) and as an approach to MT evaluation (Doherty et al. 2010). As for perceived cognitive effort, a 9-point scale was borrowed from the Educational Psychology literature. This scale ranges from 'very, very low mental effort' (1), up to 'very, very high mental effort' (9) and was originally proposed by Paas (1992) as an adaptation of previous research – having since been used in a number of projects (e.g. Roodenrys et al. 2012; Tabbers et al. 2004; Paas and Van Merriënboer 1994). Internal points in the scale were left unlabelled.

Gaze data was collected with a non-invasive remote Tobii 120Hz eye tracker (Tobii X120) via Tobii Studio (v3.1), with the Tobii VT-I fixation filter (Tobii Technology 2012). The filter was set to discard fixations below 100ms[15] – a threshold adopted in the EYE-to-IT project[16], and in previous research (Hvelplund 2011; Doherty et al. 2010).

The central rectangle in PET's interface – the ST-TT sentence area (see Fig. 2) – was established as an area of interest (AOI), which enables the automatic processing of fixations landing on this area of the screen. Further AOIs have been separately placed around the ST and TT for a computation of ST- and TT-specific eye-tracking measures. Enough space was left between the texts to allow for any spurious drifts caused by the equipment. Data for each sentence was obtained by using Tobii Studio to annotate the beginning and end of each sentence's PE process in the screen video. Eye-tracking measures were then automatically extracted with Tobii Studio for the intervals in-between.

Fixation time on screen and average fixation duration were used to estimate the quality of eye-tracking data (Rayner 1998; O'Brien 2011; Hvelplund 2011). Fixation time on screen was checked for each sentence. Based on previous research it was then checked if average fixation duration was at least 200ms for each eye-tracking session – i.e. each text. Fixation time on screen was abnormally low in two data points (less than half a second for machine-translated sentences with 8 and 25 words, respectively), denoting an inaccuracy of the equipment. These data points (0.3% of the data) were not considered for further analysis based on eye tracking. Data from all participants complied with the minimum average fixation condition.[17]

Some degree of correlation could be observed between the three response-variables used as measures of cognitive effort.[18] It is true that in situations of this kind multivariate modelling (i.e. a single model including various outcomes) could have been a more appropriate alternative to avoid redundancy and an inflated risk of type I errors due to multiple measurements. However, three separate univariate models were fit here as it seemed desirable to exploit ranks in the ordinal scale in a separate ordinal model rather than treat it as a linear variable or make use of more complex non-parametric multivariate approaches. As for the two linear outcomes,

[15] Other fixation filter settings were kept at Tobii's default values.

[16] http://cordis.europa.eu/fp7/ict/fet-open/portfolio-eyetoit_en.html

[17] One of P12's post-editing sessions only complied with this condition after excluding one of such bad-quality data points (avg. fixation duration for this session being 197ms prior to the exclusion). This session was maintained nonetheless as its removal from the data did not change the nature of results. Excluding individual data points below 200ms did not change results either and was not deemed necessary as all sessions averaged at a minimum of 200ms.

[18] Pearson's *r*: perceived cognitive effort – Fixation count: 0.62; perceived cognitive effort – avg. fixation duration: 0.43; avg. fixation duration – fixation count: 0.32.

their correlation was relatively weak (Pearson's *r* 0.32)[19], in which case the added power of multivariate modelling could be regarded as negligible (Snijders and Bosker 1999:201). Regarding the use of average fixation duration, as pointed out by Doherty et al. (2010), this measure is normally deemed capable of capturing a different aspect of cognitive load (Van Gog et al. 2009:328), which seems to denote a theoretical motivation for adopting this measure alongside the other ones considered here.

3.6 Assessing post-edited quality

Though an in-depth analysis of post-edited quality is beyond the scope of this paper, quality was assessed as a way of checking data validity. Two English native speakers assessed post-edited sentences and the raw MT output with 1-5 (low-high) scales for fluency and adequacy proposed in previous research (LDC 2005). Adequacy stands for how much of the information expressed in a reference translation is also present in a hypothesis translation. Fluency measures linguistic quality.[20] The scales were discussed with the judges beforehand and they had an opportunity to ask questions. The judges did not have a high level of proficiency in French and they were not professional translators, but they were presented with human reference translations included in the *newstest2013* dataset and instructed to use the references as a gold standard. For the assessment, the original order of each text was maintained whilst randomly combining target versions. The overall order in which texts were presented was then counterbalanced between the two judges. They had no access to information on translation producers – if they were raw MT or post-edited versions. The judges typed in scores for each sentence using a spreadsheet, as shown in Fig. 3.

Fig. 3 Assessment of post-edited quality

ID	Adequacy	Fluency			French+Reference+Translation
					Passer le test ou non?
				Take the test or not?	
14			**To take the test or not?**		
					Nous avons demandé l'avis de deux spécialistes.
				We asked two specialists for their opinion.	
15			**We have asked the opinion of two specialists.**		

Time restrictions and the difficulty in recruiting evaluators with enough availability have prevented the use of a larger number of judges in the present study. It was deemed desirable that all sentences (raw MT plus post-edited versions) be seen in the assessment. This in turn required the task to be carried out over several sittings. In response to these limitations, however, the final product is also evaluated automatically by scoring participants' post-edited text with Meteor based on the same references used to evaluate the MT output. A bilingual check on aligned FR-EN documents is also carried out with the CheckMate tool[21], which checks for issues such as 'white spaces differences', 'unexpected patterns', 'suspect patterns', etc. In addition, it is worth noting that the fact that the human evaluation was blind arguably increases the validity of a comparison between raw MT and post-edited versions. While it is true that randomly combining versions from different translators has the downside of including in the same text potentially conflicting choices of cohesive markers between sentences, scrambling target versions was deemed desirable as it may lessen order effects, whereby judges may be influenced by sequences of translations with the same level of quality. Minimizing this effect is particularly important if a blind evaluation of raw MT output is of interest, since sequences of translations with machine-like errors in the same text would most likely influence the assessment, potentially leading judges to mark down machine-translated sentences. Results on post-edited quality are reported in section 4.2.

3.7 Linguistic and psycholinguistic features

As regards the ST, further to sentence-level POS features (all divided by the number of words in each sentence) obtained by parsing the texts, sentence-level lexical frequency, type-token ratio (TTR), and lexical density (ratio

[19] Total fixation duration was also considered, but, here, a high correlation with fixation count was observed (Pearson's *r*: 0.98), arguably rendering redundant the use of both measures.

[20] The adequacy scale is as follows: 5 = All, 4 = Most, 3 = Much, 2 = Little, 1 = None. Values for fluency correspond to 5 = Flawless English, 4 = Good English, 3 = Non-native English, 2 = Disfluent English, 1 = Incomprehensible.

[21] `http://www.opentag.com/okapi/wiki/index.php?title=CheckMate`

of content words over all words) are also tested. Lexical frequency is considered in terms of the percentage of words belonging to a list of the 1K most frequent words in French, and has been used as a potential index of translating difficulty in previous research (see section 3.1), also being known to influence eye movement behaviour in reading (Rayner 1998). As for TTR and lexical density, TTR may indicate effects posed by sentence-level repetition, while lexical density may index the impact exerted by the amount of information conveyed by the ST, with higher density arguably being expected to require more processing capacity, hence more cognitive effort.

Psycholinguistic indices obtained with version 3.0 of the Coh-Metrix automatic text analysis tool[22] (Graesser et al. 2004; Graesser and McNamara 2011) are also tested – these indices correspond here to the MT output. A complete list of the Coh-Metrix indices selected for analysis is provided below.

WRDPOLc – Average polysemy for content words, based on WordNet (Miller 1995)[23]
WRDHYPnv – Mean hypernymy of nouns and verbs, measured as the number of hierarchical levels in associations maintained with other words with broader meanings in WordNet, i.e. how specific words are.
WRDAOAc – Age of acquisition of content words, i.e. the age from which speakers of English are normally expected to have the word in their vocabulary.
WRDFAMc – Familiarity score of content words, i.e. how familiar the words sound to an adult.
WRDCNCc – Index of word concreteness.
WRDIMGc – Index of word imagability, i.e. how easy it is to construct a mental image of the word.
WRDMEAc – Word meaningfulness ratings (Toglia and Battig 1978).

Even though only linguistic features based on the ST are analysed in the majority of previous research, both practical and theoretical reasons have motivated the choice of also using the English translations for this purpose here. First, to the best knowledge of the present author tools capable of providing advanced psycholinguistic indices such as the ones outlined above are currently not available for French. Second, though ST features may be mirrored in the MT, it also seemed interesting to test psycholinguistic features based directly on the machine output as, in this context, these measures could potentially denote direct indices of the complexity of information conveyed in the translation, as opposed to the ST.

To avoid over-fitting and multicollinearity problems in the statistical analysis – i.e. too many or correlated predictors – Coh-Metrix indices plus ST lexical frequency, lexical density and TTR were treated with Principal Components Analysis[24] (PCA) prior to entering the regression models. PCA is a statistical technique that takes a set of *n* variables as input, and provides *n* orthogonal (i.e. uncorrelated) principal components (PCs) as output. In addition to sentence-level POS features, six PCs, accounting for 90% of the variance in the original features altogether, and at least 5% each (Baayen 2008), were considered for further analysis.

4 Results and discussion

4.1 Textual features and cognitive effort

When inspecting the connection between Meteor and average fixation duration, fixation count, and perceived cognitive effort, an overall negative relationship can be observed, consistent with the Meteor scale, where higher values represent stronger matches with a reference, hence expected to pose less effort.

The relationship between Meteor and measures of cognitive effort is plotted in Fig. 4 and Fig. 5. Fig. 4 shows average fixation duration per sentence. As fixation count will inevitably vary with sentence length alone, it is presented in the graph per character in each source sentence, having also been log-transformed, as it did not follow a normal distribution. Both measures were z-standardised – i.e. by subtracting the mean and dividing by one standard deviation. As can be seen, Meteor seems to have a steeper negative correlation with eye-tracking measures for values below 0.6. A considerable degree of variation can be noticed for translations with high Meteor scores. Results obtained with the perceived cognitive effort scale, in Fig. 5, show a wide range of Meteor values for sentences rated as posing ‘very, very high mental effort’ (9). However, this rank was only used by participants on six occasions, which most likely explains the wide error bar.

[22] `http://cohmetrix.memphis.edu/cohmetrixpr/index.html`
[23] `http://wordnet.princeton.edu/`
[24] Using the `prcomp` R function with options `scale` and `center`.

Fig. 4 Meteor (x-axis) and standardised Avg. Fix. Duration (y-axis) per sentence (left) and standardised Log Fixation Count (y-axis) per character (right) with loess line and 95% confidence intervals

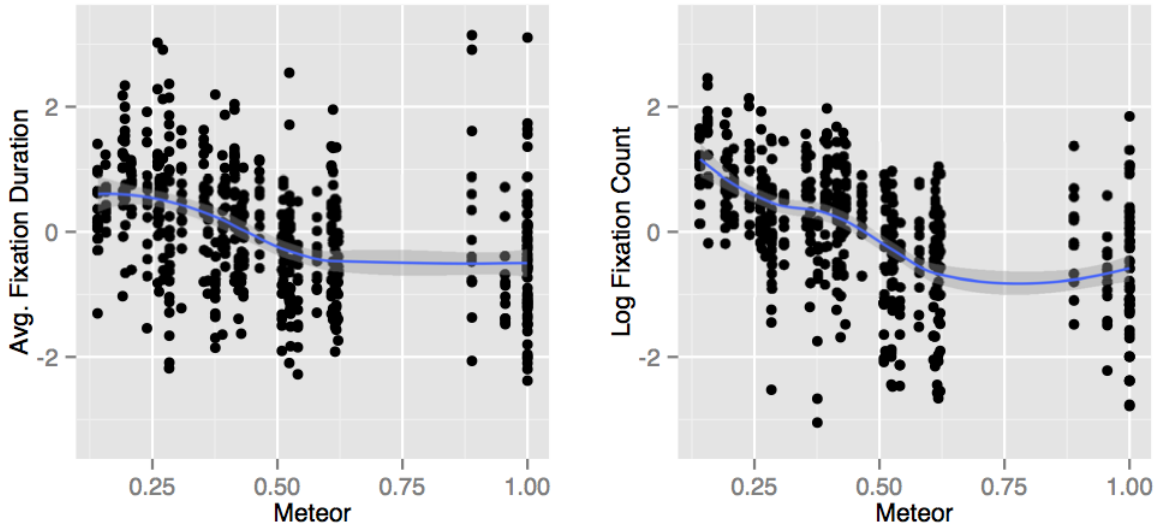
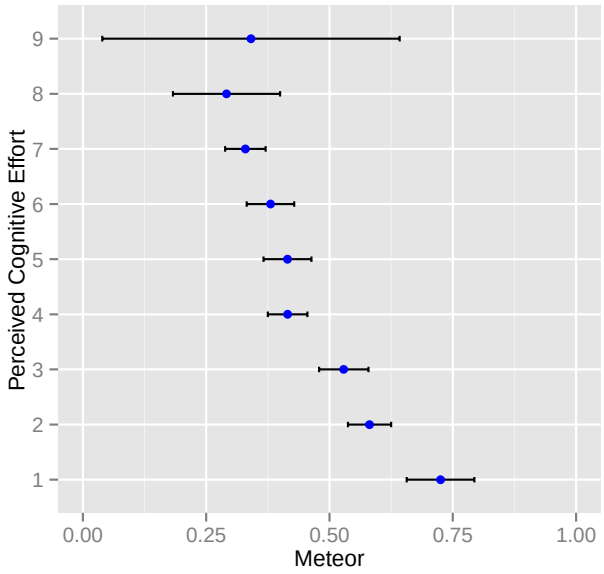


Fig. 5 Meteor (x-axis) and perceived cognitive effort (y-axis) – dots are average Meteor values per rank with 95% confidence-interval bars



To further analyse the effect of Meteor on cognitive effort whilst taking into account potential influencing factors stemming from textual features and participants' individual characteristics, two linear regression models were fit[25] with average fixation duration and log-transformed fixation count as outcome variables. Ranks of perceived cognitive effort were exploited in an ordinal model.[26] Covariates included in all three models consist of per-participant French and WMC scores, per-sentence features described in section 3.7 and source-sentence length (in characters). Regression analysis was carried out with mixed-effects models, which allow variables to be arbitrarily established as random or fixed effects. To increase confidence that any effects observed can be generalised to the wider population it is important to make sure that results are not

[25] With the `lme4` R package (Bates et al. 2013).
[26] Fit with the `ordinal` R package (Christensen 2010).

constrained to the materials or subjects sampled. As a way of tentatively tackling this issue, participants and sentences are treated as random effects in the present study.[27]

For simplicity, only one two-way interaction justified by previous research was allowed in the models. Informally, interactions take place when the impact of a given predictor on the outcome variable is moderated by other predictors. Previous research demonstrates that short sentences tend to be penalised by AEM scores (O'Brien 2011), so it was tested here if the performance of Meteor as a predictor of cognitive effort varies depending on source-sentence length.

All predictors were z-standardized prior to entering the models. Initial comprehensive models with all variables were reduced by stepwise backwards elimination (Balling and Baayen 2008:1170) until only significant effects ($p < 0.05$) remained. A potential non−linear relationship between Meteor and cognitive effort was also considered by including a quadratic term of Meteor in the models, but this effect was not significant. It was also inspected if a more complex random effects structure was required. Random slopes of Meteor by participant were justified in both linear models. By-participant random slopes of PC3 were also justified in the model with log fixation count.[28] The inclusion of random slopes in the models allows for the accommodation of effects that affect participants in different ways. Some may be faster for sentences with high Meteor scores, for example, while others may spend more time pondering on these sentences, leading to more and/or longer fixations – a variation that plots in Fig. 4 could be suggesting. Outlying data points with standardised residuals at more than 2.5 standard deviations from zero were eliminated and the reduced models were refit (ibid. 1170). P-values for predictors in all reduced models were calculated by likelihood ratio tests. No difference was observed in significance bands in comparison with results obtained with `summary`[29] functions for the ordinal or linear models, so only values based on the latter are reported. Results are presented in Table 3.

Table 3 Significant effects in three mixed-effects models: ••• $p < 0.0001$ •• $p < 0.01$ • $p < 0.05$

	Avg. Fixation Duration (in sec.)		**Log Fixation Count**		**Perceived Cog. Effort Score**	
Observations	518		524		533	
	β	**t**	**β**	**t**	**β**	**z**
Meteor	-.026	5.69 •••	-.414	6.58 •••	-1.89	6.51 •••
sentence length	-.002	.80	.364	6.24 •••	.28	1.27
prep. phrases					.46	2.19 •
PC3			.177	3.17 ••	.49	2.32 •
Meteor:sentence length	-.012	3.08 ••			-.69	2.55 •

mean avg. fixation duration: .265±.047 SD; mean fixation count: 105±86 SD; median perceived cognitive effort: 4, mode: 2

As can be seen, Meteor is highly statistically significant in the three models, with a negative correlation. Only Meteor and the interaction between Meteor and sentence length remained in the average fixation duration model. In this respect, significant interaction terms in the average fixation duration and ordinal models seem to confirm the interaction between Meteor and the length of source sentences, with Meteor being a more accurate effort predictor for longer sentences.

PC3, a principal component based on variables described in section 3.7, was significant in two models. Though it is not possible to precisely isolate the effect of original variables the PC is composed of, by inspecting the variable loadings in this PC it is noted that source TTR is the feature that contributes most to it, with a loading of -0.51. Given the positive sign of PC3 in the models, this result suggests that higher TTR might decrease cognitive effort. While repetition throughout an entire text could arguably be regarded as a facilitating effect, in the present study TTR was measured for each sentence, where repeated words in close proximity

[27] Random effects are those which are not repeatable and do not have a fixed number of levels – differently from Meteor, for example, which ranges specifically between 0 and 1 and can have repeated values across the sample.

[28] Random slopes cannot yet be implemented in the `ordinal` package, hence only included in the linear models.

[29] With the `lmerTest` R package (Kuznetsova et al. 2013) for the linear models.

might hinder the fluency of the text. An example is provided below, where the repetition of words such as 'leave' and 'years' accounted for a lack of fluency in the translation.

> *I recommend the test therefore to leave from 50 years, or to leave from 40 years if one has a direct parent who has dj had a cancer of the prostate.*

Though percentage of nouns in the ST is reported in Green et al. (2013) as a general linguistic feature for predicting PE time, results obtained here seem to indicate that this effect might not generalise to French as SL whilst having measures of cognitive effort as outcome variable. This would be consistent with the analysis of features carried out by Specia (2011) (see section 2.3), where nouns had a high correlation with PE effort for English, but not for French. As regards other POS features, a significant effect can be observed for prepositional phrases in the ordinal model. However, the effect observed is small and it loses significance if poor MT (sentences scored by both human judges with the lowest level of fluency and/or adequacy) is removed from the data, with the p-value rising to 0.07. Despite the small correlation of this feature, it would not be surprising that POS features index different phenomena depending on the SL, or its family. In the case of nouns (EN) and prepositions (FR), it is worth noting that nouns can act as direct modifiers of other nouns in English, whilst in French and other Western European Romance languages, this modification would normally require the use of prepositions.

PC3 also loses significance in the absence of low-quality machine translations in the ordinal model ($p = 0.06$), but it remains significant for Log Fixation Count. No other alterations were observed in the results upon removal of bottom-quality translations.

As for the other variables in the models, an expected positive effect is observed between the length of source sentences and number of fixations landing on the text overall.

4.2 Post-edited quality

Based on Meteor and human-assessed adequacy and fluency scores, it is possible to observe that all participants were able to improve the existing translation overall. Per-participant Meteor and average human-assessed scores are presented in Table 4, where percentages pertain to the proportion of sentences where raw MT quality was at least maintained.

Table 4 Automatic and human evaluation of post-edited quality

		Mean Adequacy 1-5 (% better or same as MT)		**Mean Fluency 1-5** (% better or same as MT)	
	Meteor 0-1	**Judge A**	**Judge B**	**Judge A**	**Judge B**
P01	0.45	4.4 (100%)	4.1 (92%)	4.4 (97%)	4 (100%)
P02	0.45	4.6 (100%)	4 (97%)	4.4 (100%)	4 (97%)
P03	0.44	4.4 (100%)	3.8 (92%)	4.4 (100%)	4.1 (92%)
P04	0.44	4.2 (97%)	3.9 (95%)	4.3 (97%)	3.8 (95%)
P05	0.47	4.7 (97%)	4.2 (92%)	4.6 (95%)	4.3 (92%)
P06	0.44	4.3 (97%)	3.9 (85%)	4.7 (100%)	4.4 (92%)
P07	0.45	4.6 (97%)	4 (90%)	4.5 (100%)	4.3 (97%)
P08	0.47	4.6 (100%)	4 (92%)	4.5 (95%)	4.1 (92%)
P09	0.47	4.7 (100%)	4.2 (95%)	4.6 (100%)	4.3 (97%)
P10	0.44	4.7 (97%)	4 (95%)	4.7 (100%)	4.1 (92%)
P11	0.45	4.2 (100%)	3.9 (97%)	4.2 (97%)	3.8 (97%)
P12	0.46	4.5 (100%)	3.8 (87%)	4.7 (100%)	4.5 (97%)
P13	0.45	4.5 (90%)	4 (85%)	4.6 (87%)	4.8 (95%)
MT	0.41	3.5	3.2	3.4	2.9

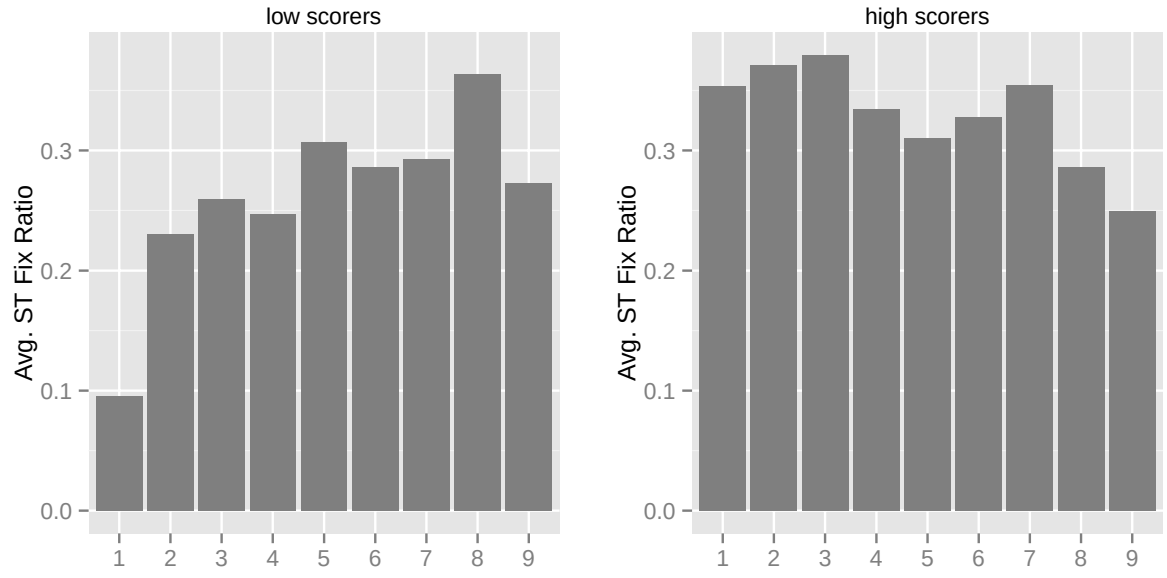
Considering evaluations from both judges, of the MT sentences rated as needing improvement – i.e. those scoring below 5 – on 491/637 occasions (77%) the post-edited version was deemed better than the MT output in adequacy, and on 607/766 occasions (79%) it was deemed better in fluency. Checks carried out with the CheckMate tool on aligned FR-EN documents flagged a single error: a case of word repetition in a sentence post-edited by P13. In total, post-edited sentences were assessed as having lower adequacy than the MT output on 50/1066 occasions (4%) and lower fluency on 36/1066 (3%). Considering this was a controlled task, where Internet searches and sentence revisits could not be allowed, these results were deemed satisfactory.

The inter-rater reliability of the human assessment was computed by calculating Cohen's weighted kappa coefficient[30] (Cohen 1968), which takes the severity of disagreements into account. Different disagreement weights are desirable in this case, since a disagreement between ranks 5 and 1 is more severe than a disagreement between ranks 4 and 5, for example. Linear weights were applied here. Recent research shows that, especially for scales with an odd number of levels, linear weights may be superior to quadratic weights (Warrens 2012), which are another traditional weighting alternative. It is worth noting, however, that no consensus seems to exist in this respect – the alternative justified by fairly recent research, and which seemed most conservative, was given preference here. Kappa values normally range from 0 to 1, corresponding to agreement due to chance and perfect agreement, respectively. The linearly weighted Kappa for adequacy was 0.39, and for fluency, 0.38. While agreement between the two judges could certainly have reached higher levels, these results seem consistent with previous research, where agreement that is only marginally above chance has been achieved in the assessment of post-edited quality (Carl et al. 2011) as well as raw MT output (Callison-Burch et al. 2007).

4.3 Fixations on the source text

To further analyse the effect of SL lexical knowledge on PE, a ratio of ST-use was computed. Based on data from all participants, 29% of all fixations landed on the ST. For P03 and P11, those with the lowest scores in the French vocabulary test, 15% and 5% of all fixations landed on the ST, respectively. While it would be expected that those with higher SL proficiency would make more use of the ST, it is interesting to notice that those with low French proficiency still looked at the ST – especially in the case of P03, who had the lowest French vocabulary result out of all participants.

Fig. 6 Relationship between average ST fix ratio (y-axis) and perceived cognitive effort (x-axis) for low French vocabulary scorers (left) and high French vocabulary scorers (right)



[30] With the *irr* R package (Gamer et al. 2012).

To analyse these figures in view of how the task was perceived, an investigation of a ratio of fixations landing on the ST – henceforth 'ST fix ratio' – and its relationship with perceived cognitive effort was carried out. Since looking at the ST would be expected to be related to participants' level of French, the possibility of a correlation between ST fix ratio and perceived cognitive effort being moderated by French knowledge is also investigated. For illustrative purposes, plots for subsets pertaining to the 6 top- and bottom-scoring participants in the French test are presented in Fig. 6. Data from the participant standing at the median of vocabulary scores was excluded from the plots in view of the odd number of participants in the study.

As can be seen, there seems to be a moderate positive trend for those in the bottom group, while a slight overall negative trend is observed for the top group. A more complex version of the ordinal model presented in Table 3 seems to confirm this association. A potential interaction of French knowledge (actual per-participant scores) and ST fix ratio (both z-standardised) was included in the model. In addition to the features presented in Table 3, the interaction between French proficiency and ST fix ratio was found to be significant ($\beta = -.48$, $z = 4.06$, $p < 0.001$). The main effect of ST fix ratio also had a significant, but small, positive association with perceived cognitive effort ($\beta = .28$, $z = 2.35$, $p < 0.05$).

Based on these results, it can be tentatively posited that ST fix ratio tends to be associated with perceived cognitive effort mostly by those whose level of SL proficiency is not very high. The perceived cognitive effort of high scorers does not seem to be related to a higher rate of ST consultation, which, in this case, may reflect mere overall checks not regarded as effortful. In fact, higher ST fix ratio was associated with less perceived cognitive effort for high scorers.

4.4 Working memory capacity and post-editing productivity

Despite not being significant in any of the mixed-effects models presented in section 4.1, group comparisons of WMC were carried out in view of PE productivity. Participants were divided into two groups split at the median of WMC scores – a 'low' WMC group including P01, P04, P05, P10, P12 and P13, and a 'high' WMC group including P03, P06, P07, P08, P09 and P11. Data from the participant standing at the median was not considered in view of the odd number of subjects in the study. The average number of words post-edited per second between the two groups was then compared for the task as a whole and for groups of machine translations split at the Meteor median. Even though PET provides an automatic measure of editing time, for consistency with eye tracking results, the duration of scenes computed with Tobii Studio was used to obtain a measure of words post-edited per second, which refers here to the number of words in the final post-edited versions, including words that were kept as in the raw MT output – a measure similar to the one used in Specia (2011). Results for the entire task are presented in Fig. 7 and for different Meteor groups in Fig. 8. Productivity was logarithmically transformed as it did not follow a normal distribution.

Fig. 7 Log average words per second (y-axis) for participants with low and high WMC (x-axis)

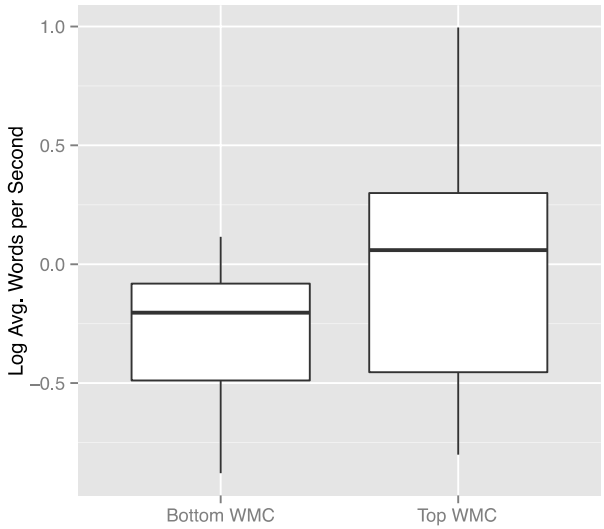
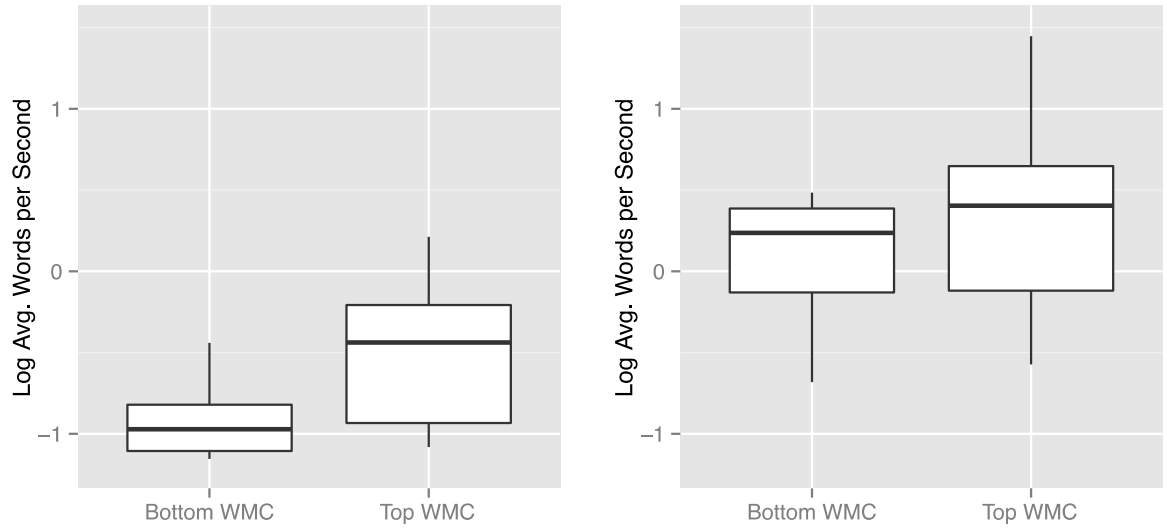


Fig. 8 Log average words per second for participants with low and high WMC for translations in the bottom Meteor quantile (left) and top Meteor quantile (right)



A pattern of more post-edited words per second can be noticed for the top WMC group. This can be observed both for the entire sample and within groups of sentences with higher and lower Meteor scores.

In the context of a mixed-effects model with productivity as response variable, however, neither a categorical dichotomised variable for high and low WMC nor actual per-participant values reach significance, denoting that further investigation is required to confirm this pattern. The same predictors tested in the models reported in section 4.1 were included in a model with log-transformed productivity (in words per second) as outcome. Results mirror those obtained for average fixation duration but with inverted signs, with more productivity being achieved for sentences with higher Meteor scores. The main effect of Meteor is significant at $p < 0.0001$ ($\beta = 0.73$, $t = 6.96$), and Meteor is also significant in an interaction with sentence length at $p < 0.01$ ($\beta = 0.31$, $t = 3.48$). In addition to treating sentences and participants as random effects, by-participant random slopes for Meteor were also justified in this model.

In regard to post-edited quality between groups with high and low WMC, results indicate that the higher productivity of those in the high WMC group might have been achieved at the expense of a slight decrease in post-edited quality. Considering scores provided by both judges for machine translations assessed as needing improvement, for the low WMC group, post-edited versions were deemed to improve raw MT output in adequacy on 232/294 occasions (78%) and in fluency on 289/354 occasions (81%). For the high WMC group, higher adequacy was observed on 222/294 occasions (75%) and higher fluency on 276/354 occasions (77%). By comparing productivity values whilst only considering sentences where there was agreement between the judges that MT was improved in adequacy and/or fluency – whilst also disregarding cases where post-edited translations were assessed by either of the judges as worse than raw MT – higher productivity is still observed for those with higher WMC.

As regards the profile of these two groups, by inspecting values in Table 2 it can be seen that experienced participants had lower WMC scores overall. Participants with more experience were also older than the remainder of the sample, which may be denoting effects of experience as well as an impact of age on WMC, an effect largely observed in previous research (see e.g. Caplan and Waters 2003). To try and account for the potential impact of these variables, a further test was carried out by comparing productivity levels of P01, P02 and P04, in a low WMC group, and P06, P08 and P09, in a high WMC group. The inclusion of participants in these subsets was carried out by taking into account those who had the most discrepant WMC scores whilst also being roughly matched by age (21-27) and previous experience (< 1,5 year), also excluding participants with extremely low scores in French vocabulary. Higher productivity was also observed here for those in the high WMC group. The difference between the two groups, however, was smaller, and stricter matching should ideally be applied to group comparisons of this kind, which would require either a larger sample or a more

controlled selection of participants. This reiterates the fact that further investigation is necessary to confirm the pattern observed here for WMC.

5 Conclusions

5.1 Research findings

The intended contributions of this paper consisted of (i) identifying predictors of cognitive effort in PE; (ii) investigating the role of the ST and SL proficiency in PE, drawing a parallel with how cognitively demanding participants perceive the task; and (iii) investigating a potential relationship between WMC and PE productivity. As regards the first item, an analysis of both individual and textual features has been carried out and, amongst the textual indices tested, Meteor presented a significant correlation with three measures of cognitive effort, showing higher predictive power for longer sentences. Prepositional phrases in the ST and a PC loaded mainly with ST sentence-level TTR are other features for which significant effects have been found. The effects observed for ST linguistic features were, however, considerably smaller, and further investigation would be necessary to confirm their impact.

A possible reason for the weaker correlation observed for these features may lie in the fact that Meteor, a measure arguably closer to capturing MT quality, was also taken into account in the statistical analysis. To the best knowledge of the present author, this is the first study where both a quality-driven feature based on the MT output and complexity-driven features based on both MT output and ST are contrastively tested in view of their impact on measures of cognitive effort. In settings of this kind, it would not be surprising that quality-driven features have better predictive power. This would be in line with recent findings from the field of QE, where AEMs based on other machine translations used as (pseudo-)references are found to contribute considerably to QE systems (Specia and Shah 2013). Information regarding textual features that correlate with PE effort is of direct interest to the estimation of MT quality, which can then be exploited for the development of PE pricing schemes. MT QE being used for this purpose is still described as 'far from being a well-established technique ready for implementation' (TAUS 2013), which justifies research initiatives such as the one undertaken here.

In regard to topic (ii), the PE process of sentences where a higher ratio of fixations landed on the ST was perceived as more effortful by those with low proficiency in the SL, and less effortful by those with high SL proficiency. In that respect, it may be that estimating translation accuracy – e.g. via QE – and establishing thresholds that allow PE to be carried out without access to the ST would open the possibility of PE being performed by those with low proficiency in the SL. Mitchell et al. (2013) have recently shown that, even when carried out by non-professionals, blind PE can lead to promising gains in fluency. Results reported here suggest that, when carried out by participants with low proficiency in the SL, a lower ratio of ST-use is also associated with less perceived cognitive effort.

As for (iii), a relationship between WMC and productivity has been observed. Subjects with higher WMC were able to process more words per second, on average, than those with low WMC. However, differences between groups were not confirmed by results obtained with mixed-effects models including other predictors. In view of the constant shift of attention between the three textual spheres involved in bilingual PE (Krings 2001) and in view of the MT error typology based on reading span proposed by Temnikova (2010), having a higher capacity of holding information in working memory would arguably be expected to have at least some impact on PE behaviour. As mentioned in section 1, it has also been shown that WMC has a positive impact on monolingual reviewing (see McCutchen 1996). Future investigation into the relationship between WMC and PE could shed further light on aspects influencing PE performance, a concept that, as previous research shows, suffers from high subject variability.

Overall, it has been shown here that a number of factors contributing to the expense of cognitive effort in PE can be identified within both texts' and subjects' characteristics. French news articles of varying complexity and readability were selected as STs. English machine translations were then post-edited by a group of participants that varied in expertise and in SL proficiency. Machine translations of varying Meteor scores were sampled and textual features of both ST and MT output were tested. It has been found that Meteor, ST prepositional phrases and sentence-level TTR have a significant association with cognitive effort in French-English PE. As for effects stemming from subjects' characteristics, those with low SL proficiency reported experiencing more cognitive effort when a higher rate of ST use could be observed. In addition, a relationship between WMC and productivity could also be noticed, but results in this respect need further

confirmation as this effect did not reach significance in the context of mixed-effects models including other covariates.

5.2 Limitations and suggestions for future work

In regard to data validity, it is true that the sequence in which tasks were carried out in the present study might have resulted in an effect of fatigue on vocabulary and WMC results, since these tasks were performed last. This sequence was adopted because, if taken before post-editing the texts, WMC and vocabulary tests would exert an impact of fatigue on data reflecting cognitive effort in PE – the main variable in the study. In this respect, counterbalancing the overall order of tasks did not seem desirable, as the nature of these tasks is quite diverse and WMC and vocabulary tests would arguably not be regarded as natural causes of fatigue in a realistic PE context. In an ideal setting, PE tasks and additional tests should perhaps be carried out on different days, or at sparse intervals of time, which in the present study was not possible in view of the limited time participants had available.

As for the design of the PE task, it was deemed that eye-tracking data would be more reliable if participants were presented with one sentence at a time, with revisits not being allowed. This certainly decreases ecological validity, as in a normal setting post-editors would have access to the entire text, being free to move backwards and forwards. Nevertheless, it is noteworthy that the methodological downside observed in a number of previous studies of having to randomise or alter the original order of sentences has been avoided in the design implemented here, which allowed participants to carry out the task in a linear fashion. Future research would expose post-editors to the entire text at one time, preferably in the environment of a CAT tool that supports a connection with eye trackers. If sentence-level features are taken into account, however, more complex data processing would be required to reliably identify fixations pertaining to individual sentences.

It is also true that the level of methodological constraint adopted here may have affected participants in different ways. Professionals may experience a higher level of frustration about not being able to revisit sentences, for example, while for students this may not conflict with previously ingrained habits. Variations of this kind have been addressed by treating participants as random effects in the statistical analysis, which increases the generalizability of results. It is also worth noting that, albeit having different levels of expertise, all participants had had previous exposure to translation, even if only in an academic context. Ideally, a larger sample, with more homogeneous expertise clusters would have been used.

Concerning the assessment of post-edited quality, a more in-depth analysis, with a larger number of judges, would have certainly provided a more thorough picture. A comprehensive analysis of post-edited quality, however, is arguably beyond the scope of this paper, which concerns mainly cognitive effort in PE. Future research could draw a more detailed parallel between these two spheres. As for the absolute human-assessed adequacy and fluency values obtained, it is noteworthy that relatively little variation can be observed in per-participant average scores (3.8 – 4.8). Since the same scales were used here to evaluate both raw MT output and post-edited translations – which is desirable if a rate of improvement is of interest –, this is arguably an expected scenario. Considering that all participants in the study were native speakers of English with some experience in translation (either as students or professionals), a high rate of post-edited translations assessed as 'Disfluent' or 'Non-native English' – levels 2 and 3 on the fluency scale – would have been surprising. The same applies for lower levels on the adequacy scale. It may be that in future research scales that are more suitable for accommodating both raw MT and post-edited versions should be deployed.

Only one version of the Meteor metric was tested in the present investigation. In future work, other metrics and different versions of Meteor – based on the technical effort score HTER (Snover et al. 2006), and on adequacy ratings – could be contrasted, perhaps also in view of their quality-estimating power based on pseudo-references, which has direct implications for the field of QE. In that respect, previous initiatives in the area have been mainly focused on PE time, HTER and subjective effort as proxies for MT quality. In view of the emergence of CAT tools that support integration with eye trackers, it would be interesting to check the performance of eye-tracking measures as a proxy for quality in QE models, which could also be carried out at a sub-sentence level. Since the sentence is usually the finest linguistic level to which Meteor is applied, a more complex investigation involving PUs or fixation units (i.e. clusters of fixations deemed to form a single unit) was not carried out here. Similarly, though the SL adopted differs from that of most previous research, only one language pair has been analysed, which, as suggested in section 4.1, limits the present findings especially with respect to effort-predicting linguistic features.

Regarding the impact of WMC on PE, though the number of subjects in the present investigation is larger than that reported in a number of previous studies, it still constitutes a relatively small sample. In that way, scaled-up studies could be carried out to further investigate this relationship, which might require larger samples to produce conclusive results.

Acknowledgements
This research has been supported by the School of Modern Languages at Newcastle University. Particular gratitude is extended to research participants as well as Dr Francis Jones, Dr Michael Jin, and Dr Ya-Yun Chen.

References

Aikawa T, Schwartz L, King R, Corston-Oliver M, Lozano C (2007) Impact of controlled language on translation quality and post-editing in a statistical machine translation environment. In: Maegaard B (ed) Proceedings of Machine Translation Summit XI, October 2007, Copenhagen, Denmark, pp 1-7

Aziz W, Castilho S, Specia L (2012) PET: A tool for post-editing and assessing machine translation. In: Calzolari N, Choukri K, Declerck T, Doğan MU, Maegaard B, Mariani J, Moreno A, Odijk J, Piperidis S (eds) Proceedings of LREC 2012 Eighth International Conference on Language Resources and Evaluation, 21-27 May 2012, Istanbul, Turkey, pp 3982-3987

Aziz W, Koponen M, Specia L (2014) Sub-sentence level analysis of machine translation post-editing effort. In: O'Brien S, Winther Balling L, Carl M, Simard M, Specia L (eds) Post-editing of machine translation: Processes and applications. Cambridge Scholars Publishing, Newcastle upon Tyne, pp 170-199

Baayen RH (2008) Analysing linguistic data: A practical introduction to statistics using R. Cambridge University Press, Cambridge

Baayen RH, Davidson DJ, Bates DM (2008) Mixed-effects modeling with crossed random effects for subjects and items. Journal of Memory and Language 59(4):390-412

Balling LW (2008) A brief introduction to regression designs and mixed-effects modelling by a recent convert. In: Göpferich S, Jakobsen AL, Mees IM (eds) Looking at eyes: eye tracking studies of reading and translation processing, Copenhagen Studies in Language 36, pp 175-192

Balling W, Baayen H (2008) Morphological effects in auditory word recognition: Evidence from Danish. Language and Cognitive Processes 23(7-8):1159-1190. doi:10.1080/01690960802201010

Bates D, Maechler M, Bolker B, Walker S (2013) Linear mixed-effects models using Eigen and S4. R package version 1.0-5. http://CRAN.R-project.org/package=lme4. Accessed 10 April 2014

Bernth A, Gdaniec C (2002) MTranslatability. Machine Translation 16(3):175-218

Bjornsson CH (1968) Lasbarhet [Readability]. Bokförlaget Liber, Stockholm

Blatz J, Fitzgerald E, Foster G, Gandrabur S, Goutte C, Kulesza A, Sanchis A, Ueffing N (2004) Confidence estimation for machine translation. In: Proceedings of the 20th International Conference on Computational Linguistics, 23-27 August 2004, Geneva, Switzerland, pp 315-321

Callison-Burch C, Fordyce C, Koehn P, Monz C, Schroeder J (2007) (Meta-)evaluation of machine translation. In: Proceedings of the Second Workshop on Statistical Machine Translation, 23 June 2007, Prague, Czech Republic, pp 136-158

Caplan D, Waters G (2003) The relationship between age, processing speed, working memory capacity, and language comprehension. Memory 13(3-4):403-413. doi:10.1080/09658210344000459

Carl M, Dragsted B, Elming J, Hardt D, Jakobsen AL (2011) The process of post-editing: A pilot study. In: Sharp B, Zock M, Carl M, Jakobsen AL (eds) Proceedings of the 8th International NLPCS Workshop. Special theme: Human-Machine Interaction in Translation. Copenhagen Studies in Language 41, 20-21 August 2011, Copenhagen, Denmark, pp 131-142

Carl M, Kay M (2011) Gazing and typing activities during translation: A comparative study of translation units of professional and student translators. Meta 56(4):952-975. doi:10.7202/1011262ar

Christensen RHB (2010) ordinal---Regression models for ordinal data R package version 2013.9-30 http://www.cran.r-project.org/package=ordinal/. Accessed 10 April 2014

Cohen J (1968) Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. Psychological Bulletin 70(4):213–220

De Almeida G (2013) Translating the post-editor: An investigation of post-editing changes and correlations with professional experience. Dissertation, Dublin City University

Denkowski M, Lavie A (2011) Meteor 1.3: Automatic metric for reliable optimization and evaluation of machine translation systems. In: Proceedings of the Sixth Workshop on Statistical Machine Translation, 30-31 July 2011, Edinburgh, UK, pp 85-91

DeStefano D, LeFevre J-A (2007) Cognitive load in hypertext reading: A review. Computers in Human Behavior 23(3):1616-1641. doi:10.1016/j.chb.2005.08.012

Doherty S, O'Brien S, Carl M (2010) Eye tracking as an MT evaluation technique. Machine Translation 24(1):1-13
Gamer M, Lemon J, Singh IFP (2012) irr: Various coefficients of interrater reliability and agreement. R package version 0.84. http://CRAN.R-project.org/package=irr. Accessed 25 April 2014
Graesser AC, McNamara DS, Louwerse MM, Cai Z (2004) Coh-Metrix: Analysis of text on cohesion and language. Behavior Research Methods, Instruments & Computers 36(2):193-202. doi:10.3758/BF03195564
Graesser AC, McNamara DS (2011) Computational analyses of multilevel discourse comprehension. Topics in Cognitive Science 3(2):371-398. doi:10.1111/j.1756-8765.2010.01081.x
Green S, de Marneffe M-C, Bauer J, Manning CD (2011) Multiword expression identification with tree substitution grammars: A parsing tour de force with French. In: Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, 27–31 July 2011, Edinburgh, UK, pp 725-735
Green S, Heer J, Manning CD (2013) The efficacy of human post-editing for language translation. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, 27 April - 2 May 2013, Paris, France, pp 439-448
Guerberof A (2014) The role of professional experience in post-editing from a quality and productivity perspective. In: O'Brien S, Winther Balling L, Carl M, Simard M, Specia L (eds) Post-editing of machine translation: Processes and applications. Cambridge Scholars Publishing, Newcastle upon Tyne, pp 51-76
Hamilton P (1979) Process entropy and cognitive control: Mental load in internalized thought processes. In: Moray N (ed) Mental workload: Its theory and measurement. Plenum Press, New York, pp 289-298
Holmqvist K, Nyström M, Andersson R, Dewhurst R, Jarodzka H, Van de Weijer J (2011) Eye tracking: A comprehensive guide to methods and measures. Oxford University Press, Oxford
Hvelplund KT (2011) Allocation of cognitive resources in translation: An eye-tracking and key-logging study. Dissertation, Copenhagen Business School
Jakobsen AL (2003) Effects of think aloud on translation speed, revision and segmentation. In: Alves F (ed) Triangulating translation. Perspectives in process oriented research. Benjamins Translation Library, vol 45. John Benjamins, Amsterdam, pp 69-95
Jensen KTH (2009) Indicators of text complexity. In: Göpferich S, Jakobsen AL, Mees IM (eds) Behind the mind: Methods, models and results in translation process research. Copenhagen Studies in Language 36. Samfundslitteratur, Copenhagen, pp 61-80
Jones G (2000) Compiling french word frequency lists for the VAT: A feasibility study. http://www.lextutor.ca/vp/fr/. Accessed 20 December 2013
Kandel L, Moles A (1958) Application de l'indice de Flesch à la langue française. Cahiers Etudes de Radio-Télévision 19:253-274
Koponen M (2012) Comparing human perceptions of post-editing effort with post-editing operations. In: Proceedings of the 7th Workshop on Statistical Machine Translation, 7-8 June 2012, Montreal, Canada, pp 181-190
Koponen M, Aziz W, Ramos L, Specia L (2012) Post-editing time as a measure of cognitive effort. In: O'Brien S, Simard M, Specia L (eds) Proceedings of the AMTA 2012 Workshop on Post-editing Technology and Practice (WPTP 2012), San Diego, USA, 28 October 2012, 10pp
Krings HP (2001) Repairing texts: Empirical investigations of machine translation post-editing processes. Kent State University Press, Kent, Ohio
Kuznetsova A, Brockhoff PB, Christensen R (2013) lmerTest: Tests for random and fixed effects for linear mixed effect models (lmer objects of lme4 package). R package version 2.0-0. http://CRAN.Rproject.org/package=lmerTest. Accessed 10 April 2014
Lacruz I, Shreve GM (2014) Pauses and cognitive effort in post-editing. In: O'Brien S, Winther Balling L, Carl M, Simard M, Specia L (eds) Post-editing of machine translation: Processes and applications. Cambridge Scholars Publishing, Newcastle upon Tyne, pp 246-272
LDC (2005) Linguistic data annotation specification: Assessment of fluency and adequacy in translations. Revision 1.5
Meara P, Buxton B (1987) An alternative to multiple choice vocabulary tests. Language Testing 4(2):142-154
McCutchen D (1996) A capacity theory of writing: Working memory in composition. Educational Psychology Review 8:299-325. doi:10.1007/BF01464076
Miller GA (1995) WordNet: A lexical database for English. Communications of the ACM 38(11):39-41
Mitchell L, Roturier J, O'Brien S (2013) Community-based post-editing of machine-translated content: Monolingual vs. bilingual. In: O'Brien S, Simard M, Specia L (eds) Proceedings of Machine Tranlation Summit XIV Workshop on Post-Editing Technology and Practice (WPTP2), 2 September 2013, Nice, France, pp 35-43
O'Brien S (2004) Machine translatability and post-editing effort: How do they relate. In: Translating and the Computer 26, November 2004, Aslib, London, UK

O'Brien S (2005) Methodologies for measuring the correlations between post-editing effort and machine translatability. Machine Translation 19(1):37-58. doi:10.1007/s10590-005-2467-1
O'Brien S (2006a) Pauses as indicators of cognitive effort in post-editing machine translation output. Across Languages and Cultures 7(1):1-21. doi:10.1556/Acr.7.2006.1.1
O'Brien S (2006b) Controlled language and post-editing. MultiLingual, October/November issue, pp 17-19
O'Brien S (2011) Towards predicting post-editing productivity. Machine Translation 25(3):197-215. doi:10.1007/s10590-011-9096-7
Paas F (1992) Training strategies for attaining transfer of problem-solving skill in statistics: A cognitive-load approach. Journal of Educational Psychology 84(4):429-434
Paas F, Van Merriënboer JJG (1994) Variability of worked examples and transfer of geometrical problem-solving skills: A cognitive-load approach. Journal of Educational Psychology 86(1):122-133
Paas F, Tuovinen JE, Tabbers H, Van Gerven PWM (2003) Cognitive load measurement as a means to advance cognitive load theory. Educational Psychologist 38(1):63-71
Plitt M, Masselot F (2010) A productivity test of statistical machine translation post-editing in a typical localization context. Prague Bull Math Linguist 93:7-16
Rayner K (1998) Eye movements in reading and information processing: 20 years of research. Psychological Bulletin 124(3):372
Read J (2007) Second language vocabulary assessment: Current practices and new directions. International Journal of English Studies 7(2):105-125
Redick TS, Broadway JM, Meier ME, Kuriakose PS, Unsworth N, Kane MJ, Engle RW (2012) Measuring working memory capacity with automated complex span tasks. European Journal of Psychological Assessment 28(3):164
Roodenrys K, Agostinho S, Roodenrys S, Chandler P (2012) Managing one's own cognitive load when evidence of split attention is present. Applied Cognitive Psychology 26(6):878-886. doi:10.1002/acp.2889
Sanders AF (1979) Some remarks on mental workload. In: Moray N (ed) Mental workload: Its theory and measurement. Plenum Press, New York, pp 41-77
Snijders T, Bosker R (1999) Multilevel analysis: An introduction to basic and advanced multilevel modeling. Sage Publications, California
Snover M, Dorr B, Schwartz R, Micciulla L, Makhoul J (2006) A study of translation edit rate with targeted human annotation. In: Proceedings of the 7th Conference of the Association for Machine Translation of the Americas 2006, 8-12 August 2006, Cambridge, USA, pp 223-231
Specia L (2011) Exploiting objective annotations for measuring translation post-editing effort. In: Forcada ML, Depraetere H, Vandeghinste V (eds) Proceedings of the 15th International Conference of the European Association for Machine Translation, 30-31 May 2011, Leuven, Belgium, pp 73-80
Specia L, Turchi M, Cancedda N, Dymetman M, Cristianini N (2009) Estimating the sentence-level quality of machine translation systems. In: Màrques L, Somers H (eds) Proceedings of the 13th Annual Conference of the European Association for Machine Translation, 14-15 May 2009, Barcelona, Spain, pp 28-37
Specia L, Raj D, Turchi M (2010) Machine translation evaluation versus quality estimation. Machine Translation 24(1):39-50
Specia L, Shah K (2013) Deliverable D2. 1.1 Quality estimation baseline software. http://www.qt21.eu/launchpad/content/delivered. Accessed 07 May 2014
Tabbers HK, Martens RL, Van Merriënboer JJG (2004) Multimedia instructions and cognitive load theory: Effects of modality and cueing. British Journal of Educational Psychology 74(1):71-81. doi:10.1348/000709904322848824
Tatsumi M (2009) Correlation between automatic evaluation scores, post-editing speed and some other factors. In: Proceedings of MT Summit XII The twelfth Machine Translation Summit, 26–30 August 2009, Ottawa, Canada, pp 332–339
Tatsumi M (2010) Post-editing machine translated text in a commercial setting: Observation and statistical analysis. Dissertation, Dublin City University
TAUS (2010) Machine translation post-editing guidelines. TAUS. https://evaluation.taus.net/resources/guidelines/post-editing/machine-translation-post-editing-guidelines. Accessed 11 April 2014
TAUS (2013) Pricing machine translation post-editing guidelines. TAUS. https://evaluation.taus.net/resourcesc/pricing-machine-translation-post-editing-guidelines. Accessed 18 January 2014
Temnikova I (2010) A cognitive evaluation approach for a controlled language post-editing experiment. In: Calzolari N, Choukri K, Maegaard B, Mariani J, Odijk J, Piperidis S, Rosner M, Tapias D (eds) Proceedings of LREC 2010 Seventh International Conference on Language Resources and Evaluation, 19-21 May 2010, Valetta, Malta, pp 3485-3490

Tobii Technology (2012) Determining the Tobii I-VT fixation filter's default values: Method description and results discussion. Tobii Technology. http://www.tobii.com/Global/Analysis/Training/WhitePapers/Tobii_WhitePaper_DeterminingtheTobiiI-VTFixationFilter'sDefaultValues.pdf. Accessed 20 December 2013

Toglia MP, Battig WF (1978) Handbook of semantic word norms. Lawrence Erlbaum, Hillsdale

Tyler SW, Hertel PT, McCallum MC, Hellis HC (1979) Cognitive effort and memory. Journal of Experimental Psychology: Human Learning and Memory 5(6):607-617

Underwood N, Jongejan B (2001) Translatability checker: A tool to help decide whether to use MT. In: Maegaard B (ed) Proceedings of MT Summit VIII Machine Translation in the Information Age, 18-22 September 2001, Santiago de Compostela, Spain, pp 363-368

Unsworth N, Heitz RP, Schrock JC, Engle RW (2005) An automated version of the operation span task. Behavior Research Methods 37(3):498-505

Unsworth N, Redick TS, Heitz RP, Broadway JM, Engle RW (2009) Complex working memory span tasks and higher-order cognition: A latent-variable analysis of the relationship between processing and storage. Memory 17(6):635-654

Van Gog T, Kester L, Nievelstein F, Giesbers B, Paas F (2009) Uncovering cognitive processes: Different techniques that can contribute to cognitive load research and instruction. Computers in Human Behavior 25(2):325-331. doi:10.1016/j.chb.2008.12.021

Warrens MJ (2012) Some paradoxical results for the quadratically weighted Kappa. Osychometrika 77(2):315-323. doi:10.1007/s11336-012-9258-4